On the Modeling of Rectangular Geometries in Room Acoustic Simulations

Enzo De Sena, *Member, IEEE*, Niccolò Antonello, Marc Moonen, *Fellow, IEEE*,
Toon van Waterschoot, *Member, IEEE*

***Abstract*—This paper is concerned with an acoustical phenomenon called *sweeping echo*, which manifests itself in a room impulse response as a distinctive, continuous pitch increase. In this paper, it is shown that sweeping echoes are present (although to greatly varying degrees) in all perfectly rectangular rooms. The theoretical analysis is based on the rigid-wall image solution of the wave equation. Sweeping echoes are found to be caused by the orderly time-alignment of high-order reflections arriving from directions close to the three axial directions. While sweeping echoes have been previously observed in real rooms with a geometry very similar to the rectangular model (e.g. a squash court), they are not perceived in commonly encountered rooms. Room acoustic simulators such as the image method (IM) and finite-difference time-domain (FDTD) correctly predict the presence of this phenomenon, which means that rectangular geometries should be used with caution when the objective is to model commonly encountered rooms. Small out-of-square asymmetries in the room geometry are shown to reduce the phenomenon significantly. Randomization of the image sources' position is shown to remove sweeping echoes without the need to model an asymmetrical geometry explicitly. Finally, the performance of three speech and audio processing algorithms is shown to be sensitive to strong sweeping echoes, thus highlighting the need to avoid their occurrence.**

I. Introduction

SPEECH and audio processing research studies often require a room acoustic model that is representative of common acoustical conditions. A simplified model of room geometry that is often used in the literature is the perfectly rectangular box, due to its similarity with many everyday spaces. While this model is only a coarse approximation of real-world rooms, it has a number of appealing properties–it leads to elegant theoretical derivations and simple algorithmic implementations, and it avoids the need to design a detailed model of room imperfections and objects present in the room. These properties are particularly appealing for research studies or applications where the modeling of room acoustics is not the central objective.

In [1], Kiyohara et al. observed that some real-world rooms that are similar to the rectangular model (e.g. a squash court) exhibit a phenomenon that they called *sweeping echo*. This is a different phenomenon from *flutter echo*. A hand-clap in the presence of strong sweeping echoes is perceived with a distinctive, continuous pitch increase. This motivated the term *sweeping* echo. Sweeping echoes cannot be observed directly in the room impulse response (RIR) or the transfer function and appear in the RIR's spectrogram as straight lines passing through the spectrogram's origin. Examples of spectrograms with sweeping echoes are presented in Section III.

Kiyohara et al. studied the sweeping echoes based on an interesting number theoretic approach, which is capable of predicting various characteristics of the phenomenon [1], [2]. The limitation of this analysis, however, is that it only applies to very specific scenarios, i.e. a cubic room [1], a square tunnel [2], or a rectangular tunnel where the squared ratio of the edge lengths is an integer number [2]. The analysis is further restricted by the assumption that both source and microphone are positioned in the middle of the enclosure.

The main contribution of this paper is a theoretical analysis that is independent of the room dimensions and of the position of source and microphone, showing that sweeping echoes are present (although to varying extents) in all rectangular geometries. The analysis is based on the rigid-wall image solution of the wave equation, i.e. the solution that formally replaces the physical boundaries surrounding an acoustic source with an equivalent lattice of image sources in free-field [3]. The analysis, which is presented in Section IV, shows that sweeping echoes are caused by the orderly time-alignment of high-order image sources positioned in proximity of the three axial directions. Section IV also gives an insight as to why more regular setups of room, source, and microphone give rise to stronger sweeping echoes, which is consistent with empirical observations reported in [1].

The simulations in Section III, are obtained using the image

The authors are with KU Leuven, ESAT–STADIUS, Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Toon van Waterschoot is also with KU Leuven, ESAT–ETC, Advise Lab, Kleinhoefstraat 4, 2440 Geel, Belgium.

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of (i) the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)", funded by the European Commission under Grant Agreement no. 316969, (ii) KU Leuven Research Council CoE PFV/10/002 (OPTEC), (iii) Interuniversity Attractive Poles Programme initiated by the Belgian Science Policy Office IUAP P7/19 Dynamical systems control and optimization (DYSCO) 2012-2017, (iv) and was supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO-Vlaanderen, T. van Waterschoot) and a Postdoctoral Fellowship (F+/14/045) of the KU Leuven Research Fund (E. De Sena). The scientific responsibility is assumed by its authors.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes room impulse responses corresponding to the spectrograms in figures 2a, 2b, 2c, 2d and 11a in WAV format. Contact enzo.desena@esat.kulueven.be for further questions about this work.

method (IM) [3] and finite-difference time-domain (FDTD) method [4]. These room acoustics models, which are briefly described in Section II, correctly predict the presence of sweeping echoes in perfectly rectangular rooms. Audio samples generated with different setups are made available at [5] and in the supplementary downloadable material associated to this paper. The IM samples can also be generated using the Matlab code provided in Appendix. Listening to these audio samples clearly reveals that the cases with regular setups do not correspond to commonly-experienced rooms. This means that particular caution has to be exercised when using the rectangular shape as geometric model since it can yield results that are very far from typical acoustical conditions.

The regularity of the lattice of image sources, which is at the basis of the phenomenon, reduces when real-world imperfections are included in the geometric model. Simulation results in Section V show, for instance, that small out-of-square asymmetries in the room geometry are sufficient to reduce the sweeping echo phenomenon significantly.

Modeling real-world imperfections requires significant computational resources and may be too cumbersome for studies that are not centered on room acoustics, or where a high level of accuracy is not required. Section VI shows that by randomizing the image sources' position, the sweeping echoes can be removed with little additional computational burden. Notice that randomization has been used previously in the IM to model diffuse reflections [6] or to reduce coloration [7]. In [7], it is unclear whether "coloration" refers to sweeping echoes or to a more typical stationary phenomenon. The author states that coloration is caused by the regular structure of image sources, which seems to support the former interpretation, but no further analysis or explanation is given.

Section VII presents a performance analysis of three speech and audio processing algorithms. This analysis shows that strong sweeping echoes alter the measured performance significantly, thus highlighting the need to avoid their occurrence in performance studies.

Finally, Section VIII concludes the paper and suggests directions for future work.

II. Background

A. The wave equation

The propagation of sound waves in a lossless fluid is governed by the wave equation [8], [9], which is the partial differential equation (PDE) with the form

$$\triangle p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = s \text{ on } \Omega \times \tau \ , \tag{1}$$

where Ω and τ denote the spatial and temporal domain, respectively, p is the sound pressure, s is the source term, $c = 343$ m/s is the speed of sound, and $\triangle$ denotes the Laplacian operator with respect to the spatial coordinates. The acoustical properties of the walls are described mathematically by the boundary conditions(BCs) [8], [9]:

$$\frac{\partial p}{\partial t} = -cZ_w \nabla p \cdot \mathbf{n} \text{ on } \partial\Omega \times \tau, \tag{2}$$

where $\partial\Omega$ denotes the boundary of Ω, $\mathbf{n}$ is the normal vector at the boundary, ∇ denotes the gradient operator with respect to the spatial coordinates, and Z_w is the characteristic wall impedance, which models the energy losses that occur at the walls. Together, the PDE and BCs form a boundary value problem. This problem is very cumbersome to solve in general. However, in the case of rectangular rooms with rigid walls, it admits a closed-form solution. This is discussed in the next subsection.

B. Solution of the wave equation for rectangular rooms with rigid walls

Under the assumption of rigid walls and monochromatic sound pressure with angular frequency $\omega = 2\pi f$, the boundary value problem (1)-(2) simplifies to [8], [9]:

$$\begin{aligned} &\text{PDE} \quad \triangle \hat{p} + \left(\frac{\omega}{c}\right)^2 \hat{p} = \hat{s} \text{ on } \Omega \\ &\text{BCs} \quad \nabla \hat{p} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \ , \end{aligned} \tag{3}$$

where the hat notation indicates a complex sound pressure and source. If the source is a point source at position $\mathbf{x}$, the sound pressure at observation point $\mathbf{x}'$ can be expressed using separation of variables as [8]:

$$\hat{p}(\mathbf{x}, \mathbf{x}') = \frac{c^2}{V} \sum_{m=0}^{\infty} \frac{\psi_m(\mathbf{x}')\psi_m(\mathbf{x})}{\omega_m^2 - \omega^2}, \tag{4}$$

where V is the volume of the domain, and ψ_m are the so-called eigenfunctions (or modes) of the problem. In the case of rectangular rooms, the eigenfunctions can be found analytically and are given by

$$\psi_m(\mathbf{x}) = \cos\left(\frac{m_x \pi}{L_x} x\right) \cos\left(\frac{m_y \pi}{L_y} y\right) \cos\left(\frac{m_z \pi}{L_z} z\right), \tag{5}$$

where (L_x, L_y, L_z) are the room dimensions, and, with abuse of notation, m is associated with one of the combinations of the non-negative integers m_x, m_y and m_z. Furthermore, the angular frequencies ω_m take the form

$$\omega_m = c\sqrt{\left(\frac{m_x \pi}{L_x}\right)^2 + \left(\frac{m_y \pi}{L_y}\right)^2 + \left(\frac{m_z \pi}{L_z}\right)^2}. \tag{6}$$

Equation (4), together with (5) and (6), is usually referred to as the Green's function for a rectangular room with rigid walls.

C. Image method for a rectangular room

In [3], Allen and Berkley proved that the Green's function for a rectangular room with rigid walls is equivalent to a time-domain solution based on the image method. The idea behind the image method is that the sound field due to a point source in a half-space bounded by a rigid wall is equal to the sound field without the wall but with an additional point source (called *image*) placed symmetrically on the opposite side of the wall. The equivalence stems from the fact that the position of the two point sources is such that the normal component of the velocity vector field vanishes at the boundary, thus satisfying the BC in (3). In rectangular rooms, each image is itself imaged, generating an infinite lattice of equivalent point

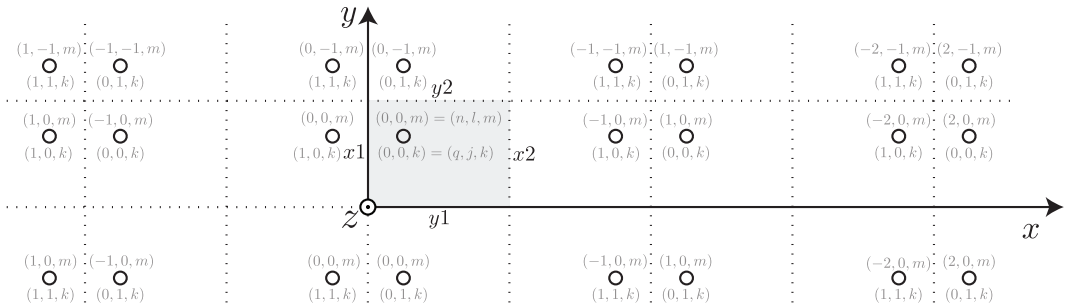


Fig. 1. Image sources with associated indices. The triplet on top of each image denote (n, l, m), while the triplet on the bottom denote (q, j, k). The labels $x1$, $x2$, $y1$ and $y2$ denote the four vertical walls.

sources as shown in Fig. 1 [3]. Allen and Berkley showed that the RIR measured at $\mathbf{x}' = (x', y', z')$ due to a point source positioned at $\mathbf{x} = (x, y, z)$ in a rectangular room with rigid walls can be expressed as[1] [3]:

$$p(\mathbf{x}, \mathbf{x}', t) = \sum_{\mathbf{p}=\mathbf{0}}^{\mathbf{1}} \sum_{\mathbf{r}=-\infty}^{+\infty} \frac{\delta\left(t - \|\mathbf{R_p} + \mathbf{R_r}\|/c\right)}{4\pi\|\mathbf{R_p} + \mathbf{R_r}\|}, \tag{7}$$

where the summations are carried out with respect to the integer vectors $\mathbf{p} = (q, j, k)$ and $\mathbf{r} = (n, l, m)$ (six summations in total, three of which only assume binary values), and where

$$\mathbf{R_p} = (x + (2q-1)x', y + (2j-1)y', z + (2k-1)z'),$$
$$\mathbf{R_r} = (2nL_x, 2lL_y, 2mL_z)\ .$$

Fig. 1 shows the indices combinations associated to some of the image sources. Allen and Berkley referred to (7) as the rigid-wall image solution of the wave equation.

While the formulation with rigid walls provides a powerful tool for theoretical derivations, it is seldom used in practical applications since it leads to a non-decaying impulse response. In the non-rigid case, however, the results are no longer mathematically exact. By assuming that image sources remain point sources, and that the wall absorption is angle-independent, the impulse response is modified to[2] [3]:

$$h(\mathbf{x}, \mathbf{x}', t) = \sum_{\mathbf{p}=\mathbf{0}}^{\mathbf{1}} \sum_{\mathbf{r}=-\infty}^{+\infty} \beta_{x1}^{|n+q|}\beta_{x2}^{|n|}\beta_{y1}^{|l+j|}\beta_{y2}^{|l|}\beta_{z1}^{|m+k|}\beta_{z2}^{|m|} \times \frac{\delta\left(t - \|\mathbf{R_p} + \mathbf{R_r}\|/c\right)}{4\pi\|\mathbf{R_p} + \mathbf{R_r}\|}, \tag{8}$$

where $\beta_{x1} \cdots \beta_{z2}$ are the wall reflection coefficients, e.g. the wall denoted by $x1$ is the wall adjacent to the origin with the normal vector parallel to the x-axis. The subscript "2", on the other hand, is associated to walls that do not pass through the origin.

D. Finite-difference time-domain method

Finite-difference time-domain (FDTD) method models room acoustics by means of time and space discretization of the wave equation. The pressure and source signals are sampled in space on a uniform grid with spacing X, and in time with temporal period T. The partial derivatives of the wave equation (1) are then approximated using central, second-order, finite differences, thus converting the PDE into a set of linear equations that can be solved iteratively.

The ratio between spatial and temporal resolution cannot be chosen arbitrarily, due to stability reasons [4]. Indeed, this ratio has to satisfy $\frac{cT}{X} \leq \lambda_{max}$, where λ_{max} depends on the specific type of discretization scheme used. This ratio, which is referred to as Courant number, is usually set with the equality in order to minimize numerical errors while conserving stability of the solution [4]:

$$c\frac{T}{X} = \lambda_{max}. \tag{9}$$

The approximations made in replacing the derivatives with finite differences introduce numerical errors that increase with frequency and cause dispersion, i.e. the phenomenon by which sound waves with different frequencies and/or directions propagate with different speeds. The dependence on the direction, in particular, makes it difficult to remove dispersion by means of post-processing operations.

The FDTD simulations in this paper will use the standard leapfrog scheme [4]. With this scheme, only the frequency range $f \in [0, 0.075 f_s]$ has a relative numerical error smaller than 2% (e.g. for $f_s = 40$ kHz, $f \in [0, 3]$ kHz) [4]. While various methodologies are available in the literature to reduce dispersion with a given sampling frequency [10], the approach used in this paper is simply to use a very high sampling frequency, at the cost of a significant computational and memory load.

III. Sweeping Echoes in Acoustic Simulations with Rectangular Geometry

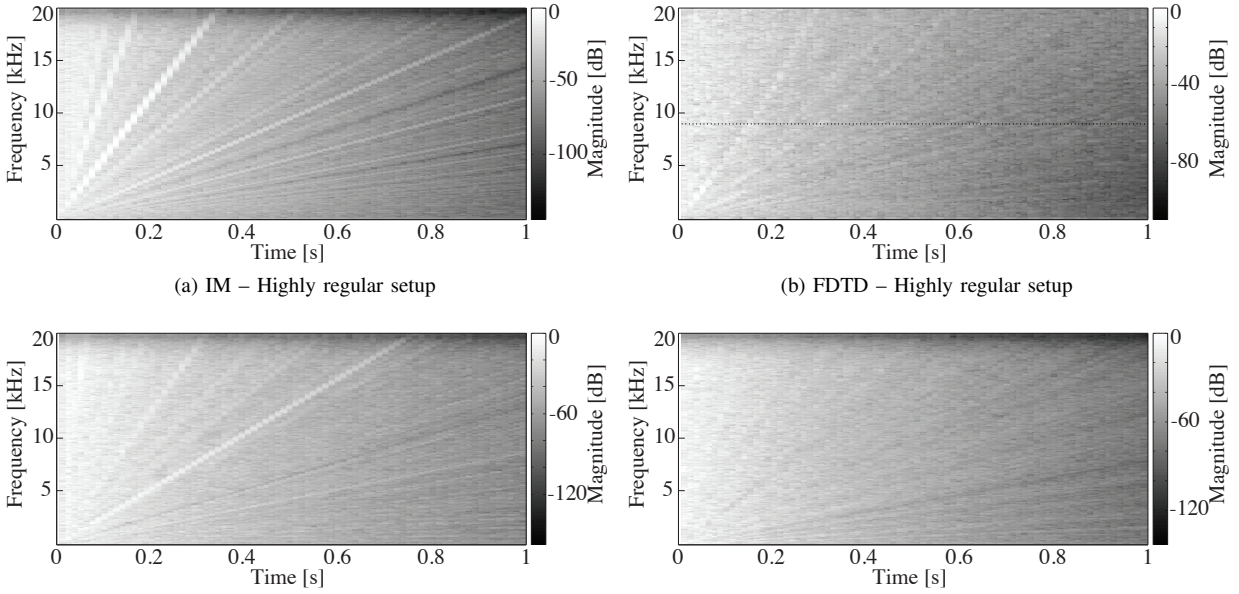
This section presents simulation results showing the sweeping echo phenomenon using the IM and FDTD method. The section concludes with the definition of a measure that quantifies the presence of sweeping echoes in the spectrogram.

A. Sweeping echoes in the IM

Fig. 2a, 2c, and 2d show the spectrogram of the RIRs produced by the IM for the three simulation setups reported in Table I with $\beta_{x1} = \cdots = \beta_{z2} = 0.93$. The sampling frequency was set to $f_s = \frac{1}{T} = 40$ kHz. The RIRs were filtered

[1]This paper uses the same notation of [3] in order to improve readability for those familiar with that paper.

[2]The indices at the exponents in this formula are summed, as opposed to [3] where they appear to be erroneously subtracted.



(a) IM – Highly regular setup

(b) FDTD – Highly regular setup

(c) IM – Example in Allen and Berkley's paper [3]

(d) IM – Irregular setup

Fig. 2. Comparison of the spectrogram of three IM-generated RIRs, 2a, 2c and 2d, and one FDTD-generated RIR, 2b, with simulation setups as in Table I. In 2b, the area below the dotted line has a relative numerical error smaller than 2%.

Figure	(L_x, L_y, L_z)	(x, y, z)	(x', y', z')	SSF
2a, 2b	$(4, 4, 4)$	$(1, 2, 2)$	$(2, 1.5, 1)$	0.5651
2c	$(8, 12, 10) \times 343/800$	$(3, 10, 4) \times 343/800$	$(5, 1, 6) \times 343/800$	0.6326
2d	$(4.1, 4.2, 4.3)$	$(1.4, 2.5, 2.6)$	$(2.7, 1.8, 1.9)$	0.9627

TABLE I
SIMULATION SETUP AND SWEEPING SPECTRUM FLATNESS (SSF) MEASURE OF RIRS IN FIG. 2.

with a 2nd-order high-pass Butterworth filter with a cutoff frequency of 50 Hz to remove non-physical behaviour at zero frequency [3]. In order to account for non-integer delays, each pulse was replaced with the impulse response of an ideal low-pass filter with cutoff frequency $f_c = 0.9\frac{f_s}{2}$, windowed with a Hann window forty samples long, as proposed by Peterson in [11]. The spectrograms were calculated over 2^{12} frequency points and used a Hamming window 25 ms long with a 50% overlap ratio. Unless stated otherwise, all remaining simulations in this paper use the above configuration with the same setup of Fig. 2a.

The spectrograms in Fig. 2a, 2c and 2d clearly show the sweeping echo phenomenon. In all three cases, it may be observed that the spectrum scales linearly with time, or, in other words, the spectrum at a given time is similar to a stretched version of the spectrum at a previous time. It may also be observed that simulation setups with a higher degree of regularity result in the phenomenon being more visible. As mentioned in the introduction, impulse responses with sweeping echoes are perceived as a distinctive audible pitch increase[3]. In the case of regular setups, the phenomenon remains clearly visible (and audible) when convolved with anechoic audio material with sharp onsets, as shown in Fig. 3 for the case of an african bongo.

[3]Audio samples of all spectrograms in Fig. 2 are available at [5] and in the supplementary downloadable material associated to this paper.

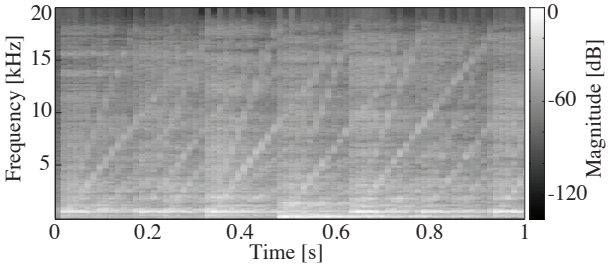


Fig. 3. Spectrogram of the convolution between the RIR in Fig. 2a and an anechoic audio sample of an african bongo. The repetitive, bright, vertical bands correspond to the sharp onsets of the audio sample.

B. Sweeping echoes in FDTD

Fig. 2b shows the spectrogram of a RIR generated using the FDTD method with the same regular setup of Fig. 2a. In this simulation, the spatial resolution is set to $X = 5$ mm. For the standard leapfrog scheme $\lambda_{max} = \sqrt{1/3}$. The sampling frequency is obtained from equation (9) as $f_s = 118$ kHz. With this sampling frequency, the standard leapfrog scheme yields a relative numerical error smaller than 2% for frequencies below 8.9 kHz. The room is excited with a physically constrained source as proposed by Sheaffer et al. in [12]. The characteristic impedance is set to $Z_w = 65$ for all walls. All remaining FDTD simulations in this paper use the above configuration.

It is clear from Fig. 2b that sweeping echoes are present

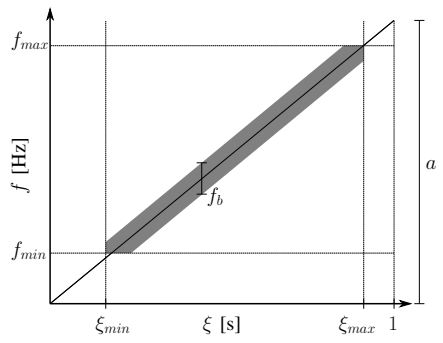


Fig. 4. Explanation of how the sweeping spectrum (SS) is calculated. The average of the normalized spectrogram is performed in the mask denoted by the gray area. The SS is the result of this averaging as a function of the slope of the line, a.

in the FDTD method too, thus confirming that this is not an artifact of the IM. The increasing pitch effect is clearly audible in the FDTD method as well. Notice also that the slope of the main sweeping echo is the same for both methods. Compared to the IM, the pattern appears slightly blurred. This is likely to be attributed to the fact that sweeping echoes, as will become clear in Section IV, are critically dependent on phase errors.

C. Sweeping echo measure

Before proceeding further, it is convenient to define a measure that quantifies how visible sweeping echoes are in the spectrogram. In order to capture the sweeping echoes' pattern, the spectrogram is averaged around a line that intersects the origin with varying slope, as shown conceptually in Fig. 4. The exact steps carried out to derive the measure are described in the remainder of this subsection.

Let the spectrogram be denoted by $|Z(\xi, 2\pi f)|^2$, where ξ is the center of the analysis window. The first step consists of normalizing the spectrogram such that the energy is identical in each time bin:

$$|\Phi(\xi, 2\pi f)|^2 = \frac{|Z(\xi, 2\pi f)|^2}{\int_{f_{min}}^{f_{max}} |Z(\xi, 2\pi f)|^2 \, df} \ , \tag{10}$$

where f_{min} and f_{max} are the minimum and maximum frequencies considered in the calculation of the proposed measure, respectively. The purpose of this normalization is to remove attenuation, which would cause lower levels of energy at lower line slopes even in absence of sweeping echoes. Let the line be described by the equation $f = a\xi$, where a is the slope of the line. The average of the normalized spectrogram is calculated as

$$\Psi(a) = \frac{1}{K} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} |\Phi(\xi_u, 2\pi f_v)| \ , \tag{11}$$

where $\mathcal{U}$ and $\mathcal{V}$ are the set of indices of the time-frequency bins that fall within the masking area shown in Fig. 4. More specifically, $\mathcal{U} = \{u | \xi_{min} \leq \xi_u \leq \xi_{max}\}$, where ξ_{min} and ξ_{max} are the minimum and maximum time bins of the mask, respectively, and $\mathcal{V} = \{v | \max(a\xi - \frac{f_b}{2}, f_{min}) \leq f_v \leq \min(a\xi_p + \frac{f_b}{2}, f_{max})\}$, where f_b is the bandwidth of the mask (see Fig. 4), and K is the number of time-frequency bins that fall within the masking area.

The function $\Psi(a)$ will be referred to as sweeping spectrum (SS) in the following. The independent variable, a, is measured in hertz per second; e.g. a strong sweeping echo centered at 10 kHz at time 1 s would be associated to $a = 10$ kHz/s.

Spectrograms with highly visible sweeping echoes have a SS with significant variation. The measure used here to quantify the variability of the SS is the spectral flatness (also known as Wiener entropy), i.e. the ratio of the geometric mean and arithmetic mean of the power spectrum [13]. This measure has values between 0 and 1, with 1 associated to the flat (white-noise-like) case, and 0 associated to the impulsive (tone-like) case. The spectral flatness of the SS, which is termed here SSF, is calculated as:

$$\text{SSF} = \frac{\sqrt[N]{\prod_{i=0}^{N-1} |\Psi(a_i)|^2}}{\frac{1}{N} \sum_{i=0}^{N-1} |\Psi(a_i)|^2} \ , \tag{12}$$

where $\Psi(a_i)$ is the SS calculated at discrete values of the slope a_i.

Table I shows the SSF measure calculated for the simulations in Fig. 2. These values were obtained using $f_{min} = 50$ Hz, $f_{max} = 0.9 f_s/2$ kHz, $\xi_{min} = 0$ s, $\xi_{max} = 0.5$ s, and $f_b = 400$ Hz. The slope values a_i were five hundred equally spaced values between 5 kHz/s and 150 kHz/s. The lower bound, 5 kHz/s, was chosen such that the line $f = a\xi$ intersects $f_{min} = 50$ Hz at $\xi = 10$ ms, which is approximately the delay of the line-of-sight component in the simulations of this paper. The higher bound, 150 kHz/s, was chosen such that all the significant sweeping echoes observed in Fig. 2 would be included. Unless stated otherwise, all SSF measures in this paper use the above configuration.

The SSF values associated to Fig. 2a, 2c and 2d are 0.56, 0.63, and 0.96, respectively. It is interesting to observe that the example in Allen and Berkley's paper [3] has a relatively low SSF value (i.e. strong sweeping echoes) even though the simulation setup was not chosen on purpose to elicit the sweeping echo phenomenon. The irregular setup has an SSF value, 0.96, which is relatively high. Notice that this value is close to typical SSF values obtained when the setup parameters are drawn from a continuous random distribution. Indeed, the SSF values of one hundred simulations with uniformly distributed room dimensions between 2 m and 8 m, and source and microphone positions uniformly distributed within the room, have an average of 0.96, with 50% of the samples falling between 0.95 and 0.97.

IV. Physical Basis of Sweeping Echoes in Perfectly Rectangular Rooms

This section gives an insight into the physical cause of sweeping echoes. The analysis is based on the rigid-wall image solution of the wave equation. The first two subsections present simplified cases with three and four walls, respectively. The analysis is then generalized to the case with six walls.

A. Simplified case with three walls

Consider the simplified case of a rectangular room where the walls denoted by $y2$, $z1$ and $z2$ are placed at a distance

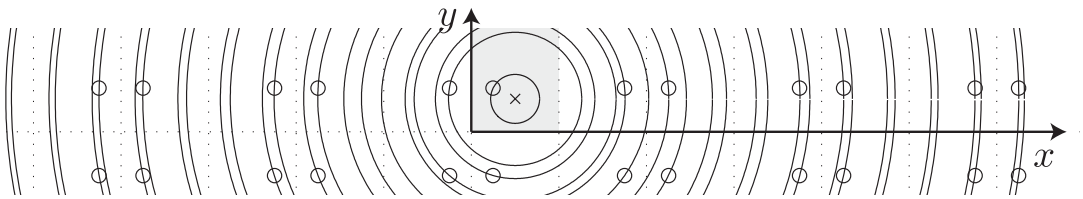


Fig. 5. Image sources for the simplified case with three walls ($x1$, $x2$, and $y1$). Circles centered at the microphone are drawn to visualize the distance of the image sources. Notice the orderly convergence between far-field pairs' distance with respect to the observation point.

large enough for the associated image sources to arrive with a significant delay. Assume also that all the walls are rigid. Fig. 5 shows the lattice of image sources associated to this problem. Notice that the regularity of image sources' positions leads to a monotonic convergence in the time arrival of far-field image pairs. In this section, it will be argued that this type of converging behaviour–which also happens in the general case–is the root cause of the sweeping echoes.

The signal formed by the converging image pairs with indices $\mathbf{p}_1 = (0, 1, 0), \mathbf{r}_1 = (n, 0, 0)$ and $\mathbf{p}_2 = (0, 0, 0), \mathbf{r}_2 = (n, 0, 0)$ for $n \geq 0$ can be written as

$$x(t) = \sum_{n=0}^{\infty} \left[\delta\left(t - \tau_n\right) + \delta\left(t - \tau_n - \Delta_n\right)\right], \qquad (13)$$

where τ_n is the delay of $\mathbf{p}_2$, and Δ_n is the difference between the time delay of $\mathbf{p}_1$ and $\mathbf{p}_2$:

$$\Delta_n = \frac{1}{c}\left(\sqrt{(2nL_x + x - x')^2 + (y + y')^2 + (z + z')^2} + \right. \\ \left. - \sqrt{(2nL_x + x - x')^2 + (y - y')^2 + (z + z')^2}\right). \qquad (14)$$

A first-order Taylor expansion of Δ_n for large n gives

$$\Delta_n \simeq \frac{yy'}{nL_x c}. \qquad (15)$$

Let τ_n be approximated by $\tau_n \simeq \frac{2nL_x}{c}$ (see Fig. 1), which implies $n \simeq \frac{\tau_n c}{2L_x}$. Substituting this expression in (15) leads to

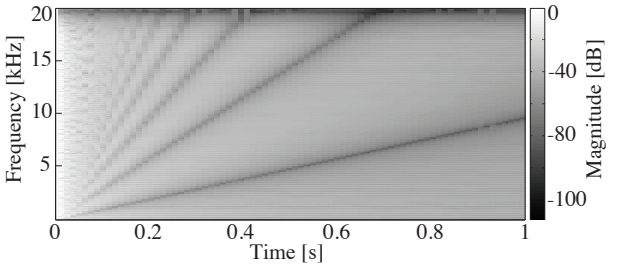
$$\Delta_n \simeq \frac{2yy'}{c^2}\frac{1}{\tau_n}. \qquad (16)$$

Consider now the windowed signal $z(\xi, t) = w(\xi, t)x(t)$, where $w(\xi, t)$ is a square window centered at ξ. Assuming that the window includes only the two impulses under consideration, the Fourier transform of $z(\xi, t)$ calculated at $\xi = \tau_n$ (i.e. when the window is centered around the delay of $\mathbf{p}_2$) is

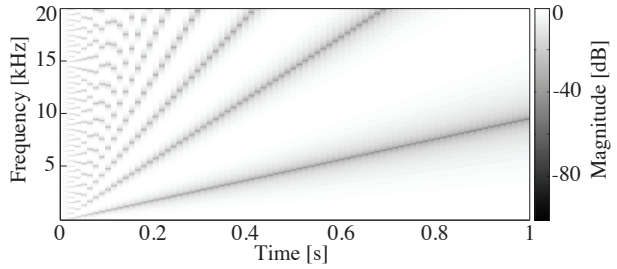
$$|Z(\xi, \omega)|^2_{\xi=\tau_n} = \left|1 + e^{-i\omega\Delta_n}\right|^2 \simeq \left|1 + e^{-i\frac{2yy'}{c^2}\frac{\omega}{\xi}}\right|^2, \qquad (17)$$

where in the last passage Δ_n was approximated as in (16).

Fig. 6 compares the actual spectrogram obtained for obtained for $x = 1, x' = 2, y = 2, y' = 1.5, L_x = 4$ with the approximation given in (17), showing a nearly exact match of the sweeping echoes behavior. This result shows that the approximations made in the derivation of (17) carry a negligible error, and that, therefore, the sweeping phenomenon is indeed caused by the progressive alignment in time of the far-field image pairs.



(a) Image solution of wave equation



(b) Closed-form approximation

Fig. 6. Simplified case with three walls. The spectrogram of the image solution of the wave equation is shown in 6a, while 6b shows the approximation derived in equation (17).

In order to gain insight into how more intricate patterns can emerge, the next subsection discusses the case where an additional wall is positioned at a finite distance from the origin.

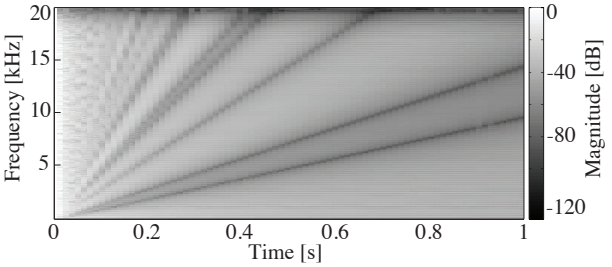
B. Simplified case with four walls

Consider the case where only the walls denoted by $y2$ and $z2$ are at a large distance from the origin. The difference from the previous example is that the xy-plane (i.e. the plane associated to the $z1$ wall) is at a close distance, and therefore the image sources are positioned on two parallel planes–one above and one below the xy-plane. Sources on each plane are still arranged as in Fig. 5. It is clear that the number of image sources with converging delay is now four. These are $\mathbf{p}_1 = (0, 0, 0)$, $\mathbf{p}_2 = (0, 1, 0)$, $\mathbf{p}_3 = (0, 0, 1)$, and $\mathbf{p}_4 = (0, 1, 1)$. Using an approach similar to the case with three walls, one obtains

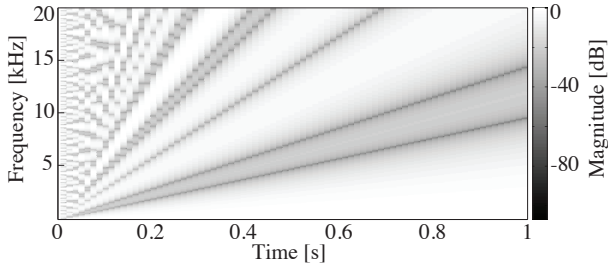
$$|Z(\xi, \omega)|^2_{\xi=\tau_n} \simeq \left|1 + e^{-i\frac{2yy'}{c^2}\frac{\omega}{\xi}} + e^{-i\frac{2xx'}{c^2}\frac{\omega}{\xi}} + e^{-i\frac{2(xx'+yy')}{c^2}\frac{\omega}{\xi}}\right|^2, \qquad (18)$$

which is plotted in Fig. 7, along with the actual spectrogram obtained for $x = 1, x' = 2, y = 2, y' = 1.5, z = 2, z' = 1, L_x = 4$.

$$\begin{aligned}\Delta\left[\mathbf{p}_1,\mathbf{r}_1;\mathbf{p}_2,\mathbf{r}_2\right] =& \simeq \frac{1}{n}\frac{1}{L_x c}\big[nL_x(\overline{q}_1-\overline{q}_2)x' + (\overline{q}_1-\overline{q}_2)xx'/2 + \\ &+ (l_1^2-l_2^2)L_y^2 + (l_1-l_2)L_y y + (l_1\overline{j}_1 - l_2\overline{j}_2)L_y y' + (\overline{j}_1-\overline{j}_2)yy'/2 \\ &+ (m_1^2-m_2^2)L_z^2 + (m_1-m_2)L_z z + (m_1\overline{k}_1 - m_2\overline{k}_2)L_z z' + (\overline{k}_1-\overline{k}_2)zz'/2\big]\ ,\end{aligned} \tag{19}$$



(a) Image solution of wave equation



(b) Closed-form approximation

Fig. 7. Simplified case with four walls. The spectrogram of the image solution of the wave equation is shown in 7a, while 7b shows the approximation derived in equation (18).

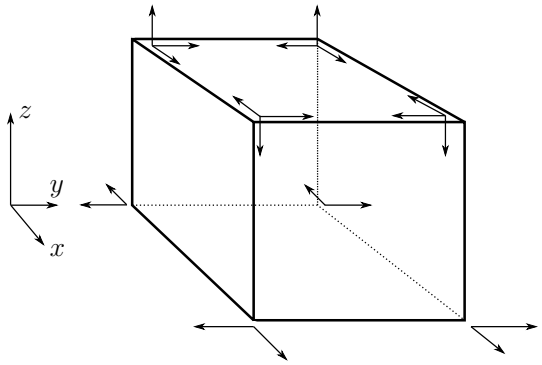


Fig. 8. Room geometry with small out-of-square distortions added to a cubic room with $L_x = L_y = L_z = 4$ m. The arrows are used to highlight the displacement of the corners. Each arrow is associated with one of the three cardinal axes. At each corner, the presence of an arrow indicates that the corner has been moved along the associated axis by δx cm in the direction of the arrow.

C. General case

The two examples above have the convenient property that only images with $l = m = 0$ are present. In the general case where all the walls are at a close distance, all the images with order $-\infty < l < \infty$ and $-\infty < m < \infty$ contribute to the patterns observed in Fig. 2, which makes it more difficult to derive approximate closed-form expressions. A first-order Taylor expansion of the difference between the delay of the images associated to $(\mathbf{p}_1, \mathbf{r}_1)$ and $(\mathbf{p}_2, \mathbf{r}_2)$ with the same order n gives the expression (19), where the over-lined indices are $\overline{q} = 2q - 1$, $\overline{j} = 2j - 1$, and $\overline{k} = 2k - 1$.

Observe in (19) that for image pairs with $\overline{q}_1 \neq \overline{q}_2$ (i.e. image pairs at opposite sides of the x-axis), Δ does not vanish for large n. All remaining pairs (i.e. image sources that lie on a plane parallel to the yz-plane) align in an orderly fashion over time for large n. Furthermore, they do so as $1/n$, or, equivalently, as $1/\xi$. This implies that, due to the scaling property of the Fourier transform ($f(\xi t) \overset{\mathfrak{F}}{\rightarrow} \frac{1}{|\xi|}F\left(i\omega/\xi\right)$ [14]), the spectrum is stretching over time. More specifically it does so linearly, i.e. proportionally to ξ, which is consistent with the sweeping phenomenon observed in Fig. 2.

Expressions similar to (19) can be obtained along the direction of the y-axis and z-axis for large l and m, respectively. Thus, the time-aligning phenomenon is present along all three axial directions. Notice that the present analysis is independent of the room dimensions, and of the position of the source and microphone. This implies that sweeping echoes are present (albeit to greatly varying extents) in all rectangular rooms.

Equation (19) also provides some insight as to why some simulation setups yield stronger sweeping echoes than others: The simulation setup of Fig. 2a is very regular, in the sense that some factors in (19) and in the corresponding expressions along the y-axis and z-axis, are identical, e.g. $L_x = L_y = L_z$, $xx' = zz'$, and others are integer multiples of each other, e.g. $L_x x = 2L_x x'$, $L_z z' = L_y y = 2L_z z$, $yy' = L_y y'$. This causes a high number of complex exponentials to sum up in phase along with their harmonics (associated to integer factors such as $l_1 - l_2$), which, in turn, yield sharper peaks.

It is worth commenting here about the similarity between sweeping echoes and the so-called *chirped echoes* [15], [16]. As opposed to sweeping echoes, a hand clap in the presence of chirped echoes is perceived with a decreasing pitch. Chirped echoes are known to occur, for instance, in front of the staircase of the El Castillo pyramid at the Maya ruins of Chichen-Itza [15]. The periodic geometry of this staircase yields a progressive temporal stretching between the reflected impulses, which, in turn, causes the decreasing pitch [16].

V. Room geometries with small out-of-square imperfections

In the previous section, it was argued that sweeping echoes are caused by the orderly time-alignment of high-order image sources. This, in turn, is due to the fact that walls opposite to each other are perfectly parallel and that adjacent walls intersect exactly at 90 degrees, which results in an extremely

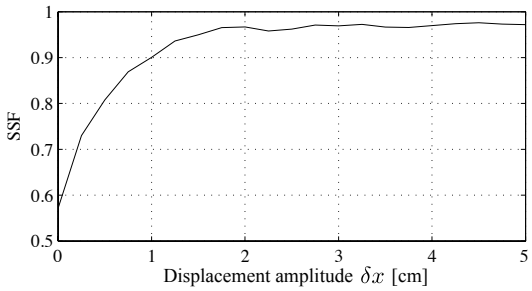
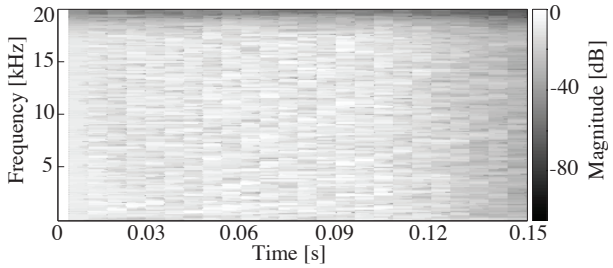
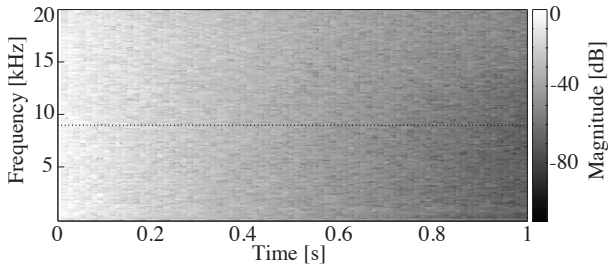


Fig. 9. Sweeping spectrum flatness (SSF) of the room geometry with small out-of-square imperfections shown in Fig. 8 as a function of the displacement amplitude δx. The plot was generated using the extended-IM. The position of source and microphone was the same of the regular setup in Fig. 2a. The case $\delta x = 0$ corresponds to a perfectly cubic room, as in Fig. 2a.



(a) Extended IM



(b) FDTD

Fig. 10. Spectrogram of the RIRs of the room geometry with small out-of-square imperfections shown in Fig. 8 with $\delta x = 2$ cm using the extended IM in 10a, and FDTD in 10b. In both cases, the position of source and microphone is the same of the regular setup in Fig. 2a. Due to the short length of the RIR in 10a, the duration of the analysis window was halved to 12.5 ms. In 10b, the area below the dotted line has a relative numerical error smaller than 2%.

regular lattice of image sources. When tilting the walls by even a small quantity, this regularity starts to break down.

To test the effect of small out-of-square asymmetries, the geometry of the room setup in Fig. 2a is distorted as described by Fig. 8. The walls are still perfectly flat, but each corner is moved by a small quantity δx in the direction indicated by the arrows. Fig. 9 shows the sweeping spectrum flatness (SSF) of this geometry as a function of the displacement amplitude, δx. The simulations were run using the IM extension to arbitrary polyhedra proposed by Borish in [17]. The Matlab toolbox EDB2 was used for this purpose [18], [19]. Specular reflections up to fourteenth order were considered, which resulted in a response around 0.15 seconds long. Orders higher than fourteen could not be considered due to significant memory and computational requirements.

Fig. 9 shows that a displacement $\delta x = 2$ cm, i.e. 0.5% of the edge length (which is considered an acceptable departure from building specifications [20]), is sufficient to increase the SSF from 0.57 to 0.97. Notice that these values should not be compared directly with the SSF values presented in Table I. Indeed, the two sets of RIRs have different durations, and the sweeping phenomenon (which is non-stationary) is measured in different parts of the spectrogram.

Fig. 10 shows the spectrograms obtained via the extended-IM and FDTD with $\delta x = 2$ cm. In FDTD, the oblique walls were approximated using a staircase implementation. It may be observed in Fig. 10 that the sweeping echo phenomenon is significantly reduced in both simulation methods.

The above results did not take into account a number of other real-world imperfections that would contribute to breaking the regularity of the image sources' lattice. These include, but are not limited to, small-scale (roughness) or large-scale (uneveness) imperfections of the walls, temperature gradients in the room, or the presence of objects. A combination of all these real-world imperfections, along with the fact that regular setups of the kind described in the previous section are unlikely in the real world, is probably the reason why sweeping echoes are not commonly experienced.

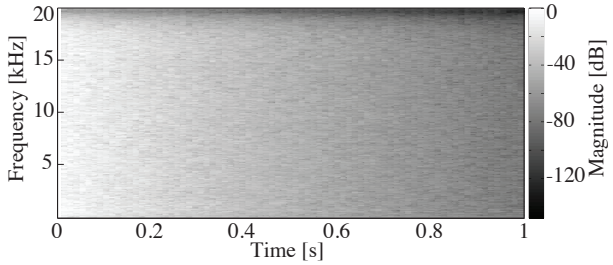
VI. Sweeping Echoes Removal from the Image Method

Modeling the real-world imperfections that break the sweeping echo phenomenon requires significant memory and computational resources. There are a number of studies and applications where such a high level of accuracy is not required, or where the computational resources are not available. Allen and Berkley's IM algorithm represents an appealing option in these cases. This motivated the analysis in this section, where it is shown how randomizing the image sources' position allows to remove the sweeping echoes with little additional computational burden. Notice that the purpose of this modification is not to model real-world imperfections or diffuse reflections as proposed in [6], but only to remove the phenomenon from the generated RIRs. In [7], Borß proposes to move each image source in a random position within the boundaries of the associated image room. As will be shown here, a much smaller random displacement is already sufficient to remove the sweeping echoes altogether.

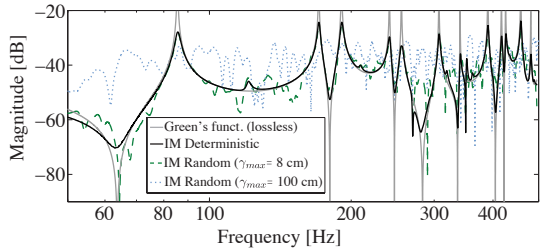
The technique used here consists of shifting each image source by a random scalar γ uniformly distributed in $|\gamma| \leq \gamma_{max}$ on the line connecting the image source and the microphone. This is equivalent to adding a random delay, γ/c, to each impulse, which makes the modification of standard IM software implementations extremely simple.

Fig. 11a shows the spectrogram of a simulation with the same setup of Fig. 2a with $\gamma_{max} = 8$ cm. It may be observed that the sweeping echoes are no longer present. The SSF of this spectrogram is 0.9945. The perceived pitch increase is also removed altogether[4]. This modification of the IM is referred to as randomized image method (RIM) in this paper.

[4]The audio sample in Fig. 11a is available at [5] and in the supplementary downloadable material associated to this paper. The same sample can also be generated using the Matlab code provided in appendix.



(a) Randomized image method



(b) Low frequency response

Fig. 11. Spectrogram of the IM (same setup of Fig. 2a) with the inclusion of a random uniformly-distributed displacement with $\gamma_{max} = 8$ cm in Fig. 11a, and low-frequency response for $\gamma_{max} = 0, 8, 100$ cm in Fig. 11b. The SSF for the spectrogram in Fig. 11a is 0.9945. Notice that the resonance peaks of the Green's function extend to infinity.

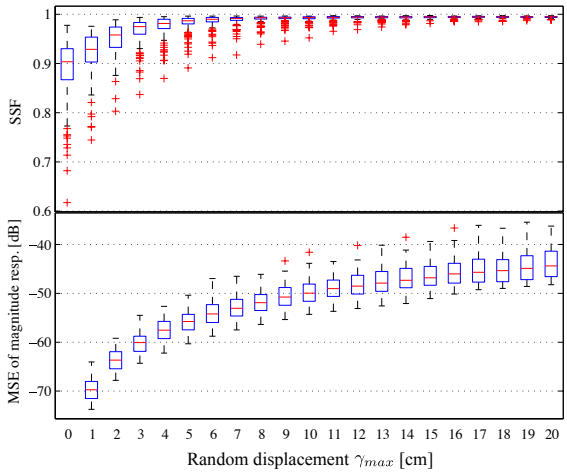


Fig. 12. Sweeping spectrum flatness (SSF) and mean square error of the low-frequency magnitude response as a function of random displacement γ. The results are calculated from a set of one hundred rooms with randomly generated dimensions, microphone position, and source position. The bottom and top extremes of the box denote the first and third quartile, respectively. The line in the middle of the box denotes the second quartile. The bars represent the extreme values, excluding the outliers. The outliers are denoted by the plus sign.

Fig. 11b shows the low-frequency response for $\gamma_{max} = 0, 8, 100$ cm. The solution of the wave equation (Green's function) for the rigid-walls (lossless) case is also overlaid, showing that there is a good match with the IM (lossy) deterministic case, i.e. $\gamma_{max} = 0$ cm. The randomized case with $\gamma_{max} = 8$ cm yields a spectrum that is largely similar to the deterministic case $\gamma_{max} = 0$ cm. Notice also that the two spectra are nearly identical around the resonance frequencies. Values of γ_{max} that are too large, on the other hand, may result in a very distorted spectrum, as shown in the figure for the case $\gamma_{max} = 100$ cm.

In order to study what is an appropriate choice for γ_{max} in more general cases, one hundred simulations were run with random room dimensions and random positions of microphone and source, the results of which are shown in Fig. 12. All random variables were chosen as discrete distributions with a small set of values so as to increase the likelihood of generating regular setups. In regular setups, in fact, sweeping echoes are more challenging to remove. The room dimensions were chosen from a discrete uniform distribution with values between 2 m and 8 m with a step of 1 m. The microphone and source were also positioned at integer distance from the walls with equal probability and were constrained to be at least 1 m away from the walls.

Fig. 12 shows the SSF and the mean squared error (MSE) of the magnitude response between 50 Hz and 500 Hz as a function of the random displacement γ. As expected, both SSF and MSE increase monotonically with γ. However, while the SSF reaches values very close to unity already at $\gamma_{max} = 8$ cm, the MSE continues to increase for larger values of γ_{max}. The SSF of $\gamma_{max} = 8$ cm, in particular, has a median value of 0.9930 (which is similar to the one in Fig. 11a), and 50% of the samples are between 0.9905 and 0.9964. The MSE of $\gamma_{max} = 8$ cm has a median value of -52 dB.

In summary, γ_{max} should be chosen based on the intended application and as a compromise between the removal of sweeping echoes and accuracy at low-frequencies. While, in general, it is suggested that the spectrogram is inspected before using a RIR, there are cases where this might be cumbersome, e.g. when the IM is used to generate a large number of RIRs with random scenarios. In these cases, $\gamma_{max} = 8$ cm appears to be a reasonable compromise.

Notice that, in case the accuracy of the magnitude response is not considered important, then other methods could be pursued for generating the late reverberant tail. A common approach with an extremely low computational footprint, for instance, is to use random noise with exponentially decaying envelope [21].

Some caution should be exercised when the room response is sought at multiple microphone positions, and when the spatial information encoded in the difference between microphone signals is important (e.g. in beamforming studies). In these cases, the image sources can be randomized relatively to a single microphone position. By using this approach, the wavefront due to each image source originates from a slightly inaccurate point, but the relative time arrivals and directions are correct across microphones.

VII. Effect of Strong Sweeping Echoes in Speech and Audio Processing Applications

This Section discusses simulation results using the IM and RIM with three different speech and audio processing applications. The purpose of this discussion is to show that the performance of different speech and audio processing algorithms is sensitive to the occurrence of strong sweeping

echoes, without assessing how the phenomenon actually affect the algorithms under consideration (which is beyond the scope of this paper).

The methodology is based on evaluating the performance of three speech and audio processing algorithms for four different sets of simulated RIRs, one of which exhibits considerably stronger sweeping echoes than the other three. It will then be shown that this particular RIR set yields significantly different results compared to the other three RIR sets. The applications considered are (i) multi-channel linear prediction of reverberant speech, (ii) objective evaluation of reverberant speech quality, and (iii) pitch estimation of monophonic reverberant music.

The four RIR sets are obtained by using the IM and RIM for simulating a cubic room of dimensions $(4, 4, 4)$ m with a fixed sound source at position $(1, 2, 2)$ m and two sets of $M = 15$ microphone positions. These two sets are selected in two distinct ways. One set of microphone positions is chosen from a uniform grid with a resolution of 0.25 m in a $(2, 2, 2)$ m box centered in the room. These microphone positions are denoted as *regular* positions, and result in highly regular setups when combined with the given room dimensions and source position. A second set of microphone positions is chosen randomly in a $(3, 3, 3)$ m box centered in the room, and are denoted as *irregular* positions. The wall reflection coefficients are all set to 0.9, and the RIRs are truncated to a length of $0.3f_s$ coefficients. For the RIM, the randomization range is set to $\gamma_{max} = 8$ cm. The SSF values for the resulting four sets of RIRs are given in Tables II and III for sampling frequencies $f_s = 8$ kHz and $f_s = 44.1$ kHz. Three out of the four RIR sets have a mean SSF value larger than 0.95, which indicates that the occurrence of sweeping echoes is weak. On the other hand, the RIR set simulated with the image method and regular microphone positions results in considerably lower SSF values, which indicates the presence of strong sweeping echoes. This can also be seen from the RIR spectrograms, which are not given here for the sake of conciseness but can easily be generated given the above simulation parameters.

Please observe that the SSF values of the two sets with different sampling frequencies should not be compared against each another. This is due to the fact that the sweeping phenomenon, which is not equally present across frequency bands, is measured in different parts of the spectrogram.

A. Multi-Channel Linear Prediction of Reverberant Speech

One of the earliest approaches to speech dereverberation is based on the hypothesis that in a linear prediction (LP) model of reverberant speech, only the LP residual is affected by reverberation while the LP model coefficients are the same as those for the clean speech signal [22]. In [23], it was shown that the latter does not hold when considering a single reverberant signal, but only holds when averaging the LP model coefficients estimated for multiple reverberant signals recorded by spatially distributed microphones in the same room.

A 12th order LP model is estimated for the phoneme /ŋ/, which is acquired from a female speaker pronouncing the Dutch word "zang" [24] and downsampled to $f_s = 8$ kHz. The estimation is performed for the clean signal, as well as for the reverberant signals obtained with each of the four RIR sets. The poles corresponding to the estimated LP model coefficients are averaged over the $M = 15$ microphone positions within each RIR set, and the mean square pole estimation error w.r.t. the poles obtained from the clean signal is calculated for each of the four RIR sets. The results are given in Table IV and indicate that the mean square pole estimation error obtained with the IM and regular microphone positions is considerably larger than for the other RIR sets.

	Microphone positions	
	Regular	Irregular
Image method	0.0242	0.0158
Randomized image method	0.0126	0.0180

TABLE IV
MEAN SQUARE POLE ESTIMATION ERRORS IN MULTI-CHANNEL SPEECH LP APPLICATION.

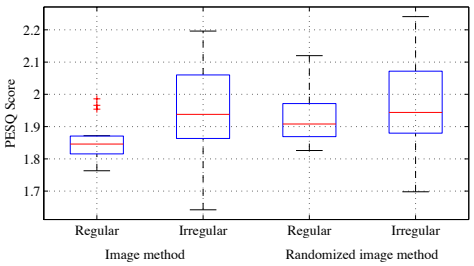


Fig. 13. PESQ score box plots for objective evaluation of reverberant speech quality.

B. Objective Evaluation of Reverberant Speech Quality

The perceptual evaluation of speech quality (PESQ) measure [25] is one of the objective measures that has been proposed in [26] for quantifying the quality of reverberant speech relative to a clean speech reference. The PESQ implementation of [27] is applied here to evaluate the speech quality of a 6.6 s female speech utterance taken from Track 49 of [28] and downsampled to $f_s = 8$ kHz. The clean speech signal is used as a reference, and the PESQ score of each of the reverberant signals obtained with the four RIR sets is calculated. Box plots of the results for the four RIR sets are shown in Fig. 13, and indicate that the PESQ score obtained with the image method and regular microphone positions is significantly lower than the scores obtained with the other RIR sets.

C. Pitch Estimation of Monophonic Reverberant Music

This section applies a pitch estimation algorithm based on approximate nonlinear least squares and maximum a posteriori estimation, included in the Multi-Pitch Estimation Toolbox of [29], to a monophonic piano signal sampled at $f_s = 44.1$ kHz. This signal is created by concatenating the C3 to C5 notes from the Steinway Grand Soft recordings in [30] so as to obtain a 60 s double chromatic scale, in which the length of each note is restricted to 2.4 s. The pitch estimation algorithm operates on

	Image method		Randomized image method	
(x', y', z')	SSF ($f_s = 8$ kHz)	SSF ($f_s = 44.1$ kHz)	SSF ($f_s = 8$ kHz)	SSF ($f_s = 44.1$ kHz)
$(1, 2, 1.50)$	0.831	0.475	0.977	0.984
$(1, 2, 1.25)$	0.920	0.630	0.979	0.991
$(1, 2, 1)$	0.821	0.448	0.980	0.986
$(1, 1.75, 1)$	0.876	0.520	0.959	0.995
$(1, 1.50, 1)$	0.833	0.442	0.955	0.988
$(1, 1.25, 1)$	0.910	0.529	0.966	0.996
$(1, 1, 1)$	0.853	0.445	0.970	0.988
$(1.25, 1, 1)$	0.861	0.512	0.980	0.991
$(1.50, 1, 1)$	0.811	0.536	0.948	0.975
$(1.75, 1, 1)$	0.840	0.502	0.971	0.993
$(2, 1, 1)$	0.879	0.582	0.910	0.948
$(2.25, 1, 1)$	0.836	0.499	0.986	0.995
$(2.50, 1, 1)$	0.805	0.529	0.923	0.972
$(2.75, 1, 1)$	0.833	0.501	0.954	0.993
$(3, 1, 1)$	0.867	0.452	0.981	0.990
mean SSF:	**0.852**	**0.507**	**0.963**	**0.986**

TABLE II
SSF VALUES FOR RIRS SIMULATED USING REGULAR MICROPHONE POSITIONS.

	Image method		Randomized image method	
(x', y', z')	SSF ($f_s = 8$ kHz)	SSF ($f_s = 44.1$ kHz)	SSF ($f_s = 8$ kHz)	SSF ($f_s = 44.1$ kHz)
$(0.819, 1.560, 2.272)$	0.961	0.971	0.991	0.994
$(2.250, 1.706, 2.337)$	0.947	0.981	0.985	0.995
$(2.305, 2.057, 0.923)$	0.968	0.983	0.980	0.995
$(0.764, 1.341, 1.624)$	0.921	0.980	0.968	0.994
$(1.584, 2.622, 2.810)$	0.968	0.984	0.978	0.997
$(1.128, 0.834, 2.352)$	0.966	0.981	0.953	0.996
$(0.863, 0.627, 3.169)$	0.947	0.977	0.980	0.994
$(2.624, 3.025, 3.213)$	0.977	0.971	0.981	0.995
$(0.646, 1.085, 1.718)$	0.973	0.973	0.980	0.993
$(3.354, 3.234, 1.694)$	0.959	0.950	0.980	0.993
$(2.921, 1.432, 1.206)$	0.948	0.947	0.947	0.993
$(2.932, 2.105, 1.044)$	0.962	0.966	0.983	0.993
$(1.713, 0.959, 1.676)$	0.966	0.977	0.961	0.993
$(0.747, 2.072, 2.569)$	0.957	0.974	0.966	0.992
$(2.642, 3.077, 3.430)$	0.946	0.973	0.983	0.996
mean SSF:	**0.958**	**0.973**	**0.975**	**0.994**

TABLE III
SSF VALUES FOR RIRS SIMULATED USING IRREGULAR MICROPHONE POSITIONS.

non-overlapping frames of 48 ms, and employs a 4096-point FFT grid size and a pitch search range of [0.01,0.1] rad/s. The mean square error, averaged over all frames, between the pitch estimated from the clean piano signal and the pitch estimated from the reverberant piano signal obtained from each of the four RIR sets, is given in Table V. Again, the result obtained with the IM and regular microphone positions stands out from the results obtained with the other RIR sets.

	Microphone positions	
	Regular	Irregular
Image method	0.335	0.274
Randomized image method	0.296	0.275

TABLE V
MEAN SQUARE PITCH ESTIMATION ERRORS FOR MONOPHONIC REVERBERANT MUSIC.

VIII. CONCLUSIONS AND FUTURE WORK

This paper was concerned with the modeling of rectangular geometries in room acoustics simulations. It was observed that perfectly rectangular geometries exhibit a phenomenon called sweeping echo. A theoretical analysis based on the rigid-walls image solution of the wave equation showed that the phenomenon is due to the progressive and monotonic convergence of the delay of high-order image sources positioned in proximity of the three axial directions. Simulation results

showed that small out-of-square asymmetries are sufficient to reduce the phenomenon significantly.

It was then observed that some room setups cause strong sweeping echoes, which are not experienced in commonly-encountered rooms. One conclusion was then that the rectangular geometry should be used with caution in cases where the objective is to model common acoustical conditions. This was highlighted by the fact that strong sweeping echoes can actually alter the performance of speech and audio processing algorithms. This also shows that previous studies employing a rectangular geometry may have been affected.

While the phenomenon was shown to be present in all rectangular rooms, it was observed that simulation parameters with a higher degree of regularity lead to stronger sweeping echoes. Surprisingly, the example setup chosen by Allen and Berkley in [3] was one of these cases. This points to the fact that manually chosen parameters (which tend to be round numbers) often result in strong sweeping echoes. Irregular setups where the room dimensions, microphone position and source positions are chosen at random from a continuous distribution, on the other hand, lead to weaker sweeping echoes.

A simple modification of the IM was proposed which consisted of delaying each image source by a small random amount. This modification was shown to remove the sweeping phenomenon altogether, while maintaining about the same computational complexity and with only a small degradation of the modal response.

In conclusion, based on the results of this paper, the following recommendation is made. In cases where the simulation setup can be chosen freely, all the parameters (i.e. room dimensions, source position and microphone position) should be drawn from a continuous random distribution. In cases where the simulation setup cannot be chosen freely and the spectrogram shows strong sweeping echoes, then, ideally, one should model the real-world imperfections that would remove the effect. Alternatively, the RIM or other types of randomization methods can be used. These methods can also be used in general cases if one wishes to have a reasonable guarantee that sweeping echoes are not occurring.

A relevant direction for future research involves measuring the perceptual impact of sweeping echoes through formal listening experiments. While it is immediately clear that strong sweeping echoes are perceivable (both in the RIR and when convolved with audio material) and do not correspond to commonly-experienced rooms, the same cannot be said for weak sweeping echoes. Informal listening tests suggest that the pitch increase in the impulse response is still noticeable for SSF values close to 0.96. However, when convolved with audio material, sweeping sounds do not appear to be perceivable. This is consistent with the results of Section VII. A relevant research direction would then be to identify the value of SSF where sweeping sounds start being clearly perceivable.

A further future work would involve investigating alternative randomization criteria. While the criterion used in Section VI has the advantage of being extremely simple, other criteria may be capable of achieving a lower error for a given increase in SSF.

Appendix: Matlab Implementation of the Randomized Image Method

Below is a self-contained Matlab implementation of the RIM method for a single source and single microphone. The only difference with the conventional IM is the instruction `+Nr*(2*rand-1)` at the beginning of the second loop.

```
function h=rim(mi, so, ro, be, Np, Nr, Tw, Fc)
% mi (microphone), so (source) and ro (room) are
% three-dimensional column vectors.
% Np: samples of the RIR.
% Nr: no. of random samples (Nr=0 for original IM).
% Tw: samples of low-pass filter, Fc: cut-off freq.
% All quantities above are in sample periods.
% be: matrix of refl. coeff. [x1,y1,z1;x2,y2,z2]
% Example 1: Fig.2a, h=rim([2;1.5;1]/343*4E4,...
% [1;2;2]/343*4E4, [4;4;4]/343*4E4,...
% 0.93.*ones(2,3), 4E4, 0, 40, 0.9);
% Example 2: Fig.11, h=rim([2;1.5;1]/343*4E4,...
% [1;2;2]/343*4E4, [4;4;4]/343*4E4,...
% 0.93.*ones(2,3), 4E4,0.08/343*4E4, 40, 0.9);
h=zeros(Np,1); ps=perm([0,1],[0,1],[0,1]);
Rps=repmat(so,[1,8])+(2.*ps-1).*repmat(mi,[1,8]);
or=floor(Np./(ro.*2))+1;
rs=perm(-or(1):or(1),-or(2):or(2),-or(3):or(3));
for i=1:size(rs'); r=rs(:, i);
 for j=1:8; p=ps(:,j); Rp=Rps(:,j);
  d=norm(2*ro.*r+Rp)+1+Nr*(2*rand-1);
  if round(d)>Np || round(d)<1; continue; end
  am=be(1,:)'.^abs(r+p).*be(2,:)'.^abs(r);
  if Tw==0; n=round(d); else
    n=(max(ceil(d-Tw/2),1):min(floor(d+Tw/2),Np))';
  end
  s=(1+cos(2*pi*(n-d)/Tw)).*sinc(Fc*(n-d))/2;
  s(isnan(s))=1; h(n)=h(n)+s*prod(am)/(4*pi*(d-1));
end; end;
function res=perm(varargin)
[res{1:nargin}]=ndgrid(varargin{1:nargin});
res=reshape(cat(nargin+1,res{:}),[],nargin)';
```

Dedication

This paper is dedicated to the memory of our friend and colleague, Nejem Huleihel (1989-2014).

References

[1] K. Kiyohara, K. Furuya, and Y. Kaneda, "Sweeping echoes perceived in a regularly shaped reverberation room," *J. Acoust. Soc. Am.*, vol. 111, no. 2, pp. 925–930, 2002.

[2] K. Kiyohara, K. Furuya, Y. Haneda, and Y. Kaneda, "Measuring sweeping echoes in rectangular cross-section reverberant fields," *Acta Acustica united with Acustica*, vol. 97, no. 2, pp. 278–283, 2011.

[3] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[4] K. Kowalczyk and M. van Walstijn, "Room acoustics simulation using 3-D compact explicit FDTD schemes," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 19, no. 1, pp. 34–46, 2011.

[5] [On-line]. Available: http://www.desena.org/sweep.

[6] B.-I. Dalenbäck, M. Kleiner, and P. Svensson, "A macroscopic view of diffuse reflection," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 793–807, 1994.

[7] C. Borß, "A VST reverberation effect plugin based on synthetic room impulse responses," in *Proc. of the 12th Intern. Cong. on Digital Audio Effects (DAFx09), Como, Italy*, 2009.

[8] F. Jacobsen and P. M. Juhl, *Fundamentals of General Linear Acoustics*. John Wiley & Sons, 2013.

[9] H. Kuttruff, *Room Acoustics*, 4th ed. SPON Press, 2000.

[10] L. Savioja and V. Välimäki, "Interpolated rectangular 3-D digital waveguide mesh algorithms with frequency warping," *IEEE Trans. on Speech and Audio Process.*, vol. 11, no. 6, pp. 783–790, 2004.

[11] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1527–1529, 1986.
[12] J. Sheaffer, M. van Walstijn, and B. Fazenda, "Physical and numerical constraints in source modeling for finite difference simulation of room acousticsa)," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 251–261, 2014.
[13] J. E. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag New York, Inc., 1982.
[14] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & Systems (2nd edition)*. Prentice-Hall, 1996.
[15] N. F. Declercq, J. Degrieck, R. Briers, and O. Leroy, "A theoretical study of special acoustic effects caused by the staircase of the El Castillo pyramid at the Maya ruins of Chichen-Itza in Mexico," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3328–3335, 2004.
[16] F. A. Bilsen, "Repetition pitch glide from the step pyramid at Chichen Itza," *J. Acoust. Soc. Am.*, vol. 120, no. 2, pp. 594–596, 2006.
[17] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.
[18] [On-line]. Available: http://www.iet.ntnu.no/∼svensson/software/.
[19] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2331–2344, 1999.
[20] *Guide to standards and tolerances*. Vistorian Building Commission, 2007.
[21] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 20, no. 5, pp. 1421–1448, 2012.
[22] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.
[23] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. 2003 Int. Workshop Acoustic Echo Noise Control (IWAENC '03)*, Kyoto, Japan, Sep. 2003, pp. 99–102.
[24] J. Wouters, W. Damman, and A. J. Bosman, "Vlaamse opname van woordenlijsten voor spraakaudiometrie," NKO – K.U.Leuven/U.Z.Leuven, 1994.
[25] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU Std. P.862, 2001.
[26] K. Kinoshita *et al.*, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. 2013 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '13)*, New Paltz, NY, USA, Oct. 2013.
[27] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, 2007.
[28] "Sound quality assessment material recordings for subjective tests," EBU SQAM CD, 2008. [Online]. Available: https://tech.ebu.ch/publications/sqamcd
[29] M. G. Christensen and A. Jakobsson, *Multi-pitch estimation*. Morgan & Claypool Publishers, 2009.
[30] F. Opolko and J. Wapnick, *McGill University Master Samples*. McGill University, 2006, DVD edition.

Enzo De Sena (S11, M'14) received the B.Sc. in 2007 and M.Sc. *cum laude* in 2009, both from the Università degli Studi di Napoli "Federico II" in Telecommunication Engineering. In 2013, he received the Ph.D. degree from King's College London in Electronic Engineering.

He is currently a Postdoctoral Research Fellow at KU Leuven. From 2012 to 2013 he was a Teaching Fellow at King's College London. From 2013 to 2015 he was a Marie Curie Fellow in the "Dereverberation and Reverberation of Audio, Music, and Speech" ITN at KU Leuven. He previously collaborated with the Network Research Lab at UCLA (2007-2009), and he was a Visiting Researcher at the Center for Computer Research in Music and Acoustics at Stanford University (2013) and at the Signal and Information Processing section at Aalborg University (2014-2015).

His current research interests include room acoustics modelling, multichannel audio systems, microphone beamforming and binaural modelling. He is a member of IEEE, EURASIP, and the Acoustical Society of America.

Niccolò Antonello received his B.Sc. in Electronic Engineering at the Università degli Studi di Padova and his M.Sc. in Acoustic Engineering at the Technical University of Denmark (DTU) in 2010 and 2012, respectively. Between September 2012 and February 2013, he was a Research Assistant at DTU focusing on suppression of loudspeaker non-linearities. He is currently pursuing the Ph.D. degree at KU Leuven as an Early Stage Researcher in the Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)".

His research interests include room acoustic simulations, dereverberation, sound reproduction, optimal control, acoustic impedance identification.

Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received journal best paper awards from the IEEE Transactions on Signal Processing (with Geert Leus and with Daniele Giacobello) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012).

He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in IEEE Signal Processing Magazine (2012-2014), and has been a member of the editorial board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, EURASIP Journal on Applied Signal Processing and EURASIP Signal Processing.

Toon van Waterschoot (S'04, M'12) received the MSc degree (2001) and the PhD degree (2009) in Electrical Engineering, both from KU Leuven, Belgium.

He is currently a tenure-track Assistant Professor at KU Leuven, Belgium. He has previously held positions as a Teaching Assistant with the Antwerp Maritime Academy, Belgium (2002), as a Research Assistant with KU Leuven, Belgium (2002-2009) and with the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT), Belgium (2003-2007), and as a Postdoctoral Research Fellow with KU Leuven, Belgium (2009-2010), Delft University of Technology, The Netherlands (2010-2011), and with the Research Foundation - Flanders (FWO), Belgium. Since 2005, he has been a Visiting Lecturer at the Advanced Learning and Research Institute of the University of Lugano (Universit della Svizzera italiana), Switzerland. He has been teaching courses related to digital signal processing, control theory, and numerical optimization. His research interests are in acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

Dr. van Waterschoot has been serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing, and as a Guest Editor for Signal Processing. He has been a Nominated Officer for the European Association for Signal Processing (EURASIP), and a Scientific Coordinator of the FP7-PEOPLE Marie Curie Initial Training Network on Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS). He has been serving as an Area Chair for Speech Processing at the European Signal Processing Conference (EUSIPCO 2010, 2013-2015), and will be the General Chair of the 60th AES Conference to be held in Leuven, Belgium, 2016. He is a member of the Audio Engineering Society, the Acoustical Society of America, EURASIP, and IEEE.